AD-A202 898

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

DTIC FILE COPY

1a. REPORT SECURITY CLASSIFICATION: UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT: Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S): AFOSR-TR- 88-1280
6a. NAME OF PERFORMING ORGANIZATION: John Hopkins University; 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION: AFOSR/NE
6c. ADDRESS (City, State and ZIP Code): Charles and 34th Street, Baltimore, MD 21218	7b. ADDRESS (City, State and ZIP Code): Bldg 410, Bolling AFB, DC 20332-6448
8a. NAME OF FUNDING/SPONSORING ORGANIZATION: AFOSR/NE; 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: AFOSR-86-0246
8c. ADDRESS (City, State and ZIP Code): Bldg 410 Bolling AFB DC	10. SOURCE OF FUNDING NOS.

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT NO.
61102F	2305	B3	

11. TITLE (Include Security Classification): Massively-Parallel Architectures for Automatic Recognition of Visual Speech Signals

12. PERSONAL AUTHOR(S): Terrence J. Sejnowski , Professor

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Yr., Mo., Day)	15. PAGE COUNT
Annual	FROM _____ TO _____		

16. SUPPLEMENTARY NOTATION

DTIC ELECTE DEC 16 1988 S D

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB. GR.	

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

During the last year significant progress has been made inthe primary objective of estimating the acoustic characteristics of speech from the visual speech signals. Neural networks have been trained on a database of vowels. The rqw images of faces, aligned and preprocessed, were used as input to these network which were trained to estimate the corresponding envelope of the acoustic spectrum. The performance of the networks was better than trained humans and was comparable with optimized pattern classifiers. Our approach avoids the problems of information loss through early categorization. The acoustic information that the network extracts from the visual signal can be used to supplement the acoustic signal in noisy environments, such as cockpits. During the next year we extend these results to diphthongs using recurrent neural networks and temporal sequences of input images. (AR)

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS ☐	UNCLASSIFIED

22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE NUMBER (Include Area Code)	22c. OFFICE SYMBOL
Giles	(202) 767-4931	NE

DD FORM 1473, 83 APR EDITION OF 1 JAN 73 IS OBSOLETE.

88 12 15 010

SECURITY CLASSIFICATION OF THIS PAGE

TO: Air Force Office of Scientific Research

FROM: Johns Hopkins University
Charles and 34th Street
Baltimore, MD 21218

PROJECT TITLE:

Massively-Parallel Architectures for Automatic Recognition of Visual Speech Signals

AWARD NUMBER: AFOSR-86-0246

PRINCIPAL INVESTIGATOR:

Terrence J. Sejnowski
Professor of Biophysics
277-42-8904

CO-PRINCIPAL INVESTIGATOR:

Moise H. Goldstein, Jr.
Professor of Electrical and Computer Engineering
435-34-6457

RESEARCH ASSISTANT:

Ben P. Yuhas
Doctoral Student in Electrical and Computer Engineering

RENEWAL PROGRESS REPORT: October 12, 1988

SUMMARY

During the last year significant progress has been made in the primary objective of estimating the acoustic characteristics of speech from the visual speech signals. Neural networks have been trained on a database of vowels. The raw images of faces, aligned and preprocessed, were used as input to these network which were trained to estimate the corresponding envelope of the acoustic spectrum. The performance of the networks was better than trained humans and was comparable with optimized pattern classifiers. Our approach avoids the problems of information loss through early categorizaton. The acoustic information that the network extracts from the visual signal can be used to supplement the acoustic signal in noisy environments, such as cockpits. During the next year we plan to extend these results to diphthongs using recurrent neural networks and temporal sequences of input images.

INTRODUCTION

Speaking produces acoustic and visual signals. When the acoustic speech signal is degraded by noise, the visual signal can provide supplemental speech information that improves speech perception (Sumby and Pollack, 1954; Ewersten and Nielsen, 1971; Erber, 1975). When the acoustic signal is totally unavailable, as with the profoundly deaf, then the visual signal can be used through lip reading. (Montgomery, 1983; Summerfield, 1979; Demorest, Bernstein and Eberhardt, 1987).

In this project, neural networks are being used to process visual speech signals in order to study the feasibility of obtaining acoustic constraints directly from the raw visual image. At present, automatic speech recognition systems rely almost exclusively on the acoustic speech signal. This contributes to the poor perform of these systems often demonstrated in noisy environments (Allen, 1985). While some effort has been made at cleaning up the acoustic input to these systems, few have attempted to use the visual information to supplement the acoustic signal.

The only speech recognition system that has extensively used the visual signals was developed by Eric Petajan of Bell Labs (Petajan, 1984, 1987). This system uses stored templates to identify sequences of lip images. On a limited vocabulary, Petajan was able to demonstrate that using the visual speech signals can significantly improve speech recognition over acoustic recognition alone.

The computational constraints imposed by serial, digital computers often makes it necessary to encode stored templates. The von Neumann bottleneck between the memory and the processors requires that the incoming images and the stored templates have a reduced dimensionality that minimizes the necessary computation. In this system, Petajan uses vector-quantization to construct a codebook that is used to translate incoming image sequences into symbol strings. The question is whether the encoding process preserves the relevant information. Petajan found that encoding the images resulted in a poorer performance, suggesting that there was a loss of information.

The neural networks is an alternative architecture characterized by many interconnected processors that perform their computation in parallel. This architecture offers new approaches to signal processing by eliminating the need to immediately encode signals into a lower dimension.

THE VISUAL AND ACOUSTIC SIGNALS OF SPEECH

In linguistics, the continuous speech signals are traditionally treated as a sequence of discrete components. *Phonemes* are the shortest acoustically distinguishing unit of a given language. For example, the words *beat* and *neat* are distinguished by the to the phonemes [b] and [n]. Similarly, *boot* and *beat* are distinguished by the phonemes [u] and [i], which are abstractions corresponding to the 'oo' and 'ea' sounds in the two words. The sounds themselves are identified *phonetically* as /u/ and /i/ to distinguish them from the linguistic abstractions [u] and [i].

The visual correlate of the phoneme is the *viseme*, The viseme is the smallest visibly distinguishing unit of a given language (Fisher, 1968). The mapping between the phonemes and visemes is generally many to one; for example, the phonemes [p],[b] and [m] are usually visibly indistinguishable and treated as a single viseme.

Speech recognition research has been largely preoccupied with trying to find a reliable method of translating the continuous acoustic signal into a corresponding phonemic sequence (Reddy, 1966, 1967). This effort has been plagued by problems in segmenting the continuous signals and also at the level of identifying those segments. Recently, however, the most successful speech recognition systems avoid this procedure all together (Jelinek, 1985). These results suggest that alternative approaches of using the visual signal should also be explored.

The acoustic speech signal that is emitted from the mouth has long been modeled as the response of the vocal tract filter to a sound source (Fant, 1960; Flanagan, 1972). In this first-order model, it is the configuration of the articulators that define the vocal tract f ilter's shape, and its corresponding resonance characteristics. These resonance characteristics are represented in the acoustic waves short-term power spectrum's amplitude envelope.

While some of the articulators are visible on the face of the speaker (e.g., the lips, teeth and sometimes the tongue), others are not. The visible articulators' contribution to the acoustic signal result in speech sounds that are much more susceptible to acoustic noise distortion than are those

contributions of the hidden articulators (Petajan, 1987). As a result, the visual speech signal tends to complement the acoustic signal. Thus, those speech sounds that are the most visibly distinct, such as /b/ and /k/, are among the first pairs to be confused when presented acoustically in the presence of noise. Similarly, those phonetic segments that are visibly indistinguishable, such as /p/,/b/ and /m/ are among the most resistant to confusion when presented acoustically (Miller and Nicely, 1955; Walden, et.al, 1977). This complementary structure serves as the basis on which the two signals can interact to improve the perception of speech in noise.

RESEARCH OBJECTIVES

The focus of this research is to study the feasibility of using the visual speech signal to define acoustic constraints. The approach has been to try and estimate the short-term power spectral envelope of the acoustic signal from the corresponding visual signals on the face of the speaker. The transfer function of the vocal tract can then be described from this spectral envelope.

Estimates of the transfer function are obtained using a variety of neural networks architectures. These estimates are then evaluated and compared with results from other estimation techniques.

The neural network architectures that we are working with are simulated on an ANALOGIC AP5000 array-processor. These architectures are in being implemented in special parallel hardware, and eventually should be readily available.

NEURAL NETWORK BACKGROUND

Neural networks compute using many interconnected processors that individually perform a simple transformation of their summed inputs. The connections between processors have weights associated with them and signals traveling along those connections are multiplied by those weights. The network's response to a particular input is based on the exchange of signals between processors across these weighted connections. A particular response can be *programmed* by specifying the connections and their associated weights. An introduction to neural networks can be found in April 1987's IEEE ASSP Magazine (Lippman, 1987; Rumelhart, McClelland, and the PDP Research Group, 1986).

By defining the correct weights, these networks can be constructed to solve a variety of problems (Hopfield and Tank, 1986; Marr and Poggio, 1976). However, until recently defining these weights was an arduous task that required an *a priori* solution to the underlying problem. Now algorithms exists that iteratively adjust these weights given a set examples (Pineda, 1987; Rumelhart, et.al.,1985), The ability to automatically program these networks has resulted in a flurry of experimental work aimed at demonstrating the computational power of this architecture. Using these algorithms, networks have already demonstrated their ability to find solutions to a variety of problems (Lippman, 1987;Sejnowski and Gorman, 1988; Sejnowski and Rosenberg, 1987; Pragar et.al., 1986;).

One of the areas neural networks is strongest is in solving ill-posed problems, where a solution may not exist or may not be unique (Poggio, et al., 1985). Estimating acoustic structure from visual speech signals is such an ill-posed problems. The visual signals provide only a partial description of the vocal tract transfer function, and that which is described is ambiguous. For a given visual signal there are many possible corresponding acoustic structures. What we want is a *good* estimate based upon known examples.

THE APPROACH

A variety neural network architectures were trained to approximate the acoustic signal's power spectrum envelope given the corresponding visual signal as input. Since the system was given single isolated video images, which correspond to 33ms of speech, it was necessary to choose data from vowels and diphthongs, which are relatively steady state over periods of 50 to 80 ms.

The input signal structure was chosen to exploit the distributed representations that neural networks allow. The video input signals are extracted from recordings of full-faced, well-illuminated speakers preserved on laser disc (Bernstein and Eberhardt, 1987). Software was written to automatically define a box centered about the mouth and extract that portion of the image. The computational

load of simulating a neural network on a serial machine made it necessary to further sub-sample these images, reducing the input to 500 pixels. It is important to emphasize that the particular encoding we have chosen is an artifact of our simulation. Once these networks are implemented in hardware, visual images can be received in parallel across arrays of sensors and then processed in parallel without sampling.

Corresponding to each video frame is 33ms of acoustic speech. When the visual signal is presented to the network, the network is asked to produce the amplitude envelope of the 256 point short-term power spectrum (STPS) of the corresponding acoustic signal. During the training process, the network is presented with both the input and output structures. To obtain the spectral envelope, the cepstrum was taken of the STPS and the low-pass lifted below T/4, where T is the original length of the input signal. Taking the inverse cepstrum of the resulting data provided a smooth envelope of the original power spectrum that could be sampled down to 32 points.

The networks were trained on 119 images of vowels and diphthongs. The training period usually involved 450 presentations of each image in the training set. The weights were updated after all images were presented. For a given network architecture, the particular weights to be evaluated will be those that give the network its best performance for data not in the training set.

EVALUATION

The performance of the network is evaluated on two criteria. The first criterion is the weighted squared error between the spectral envelope produced by the network and the true envelope. The weighting places greater emphasis on the peaks of the spectral envelope than on the valleys. This is accomplished by multiplying the squared difference of two points by the amplitude of the greater value of those two points. For a given input image, I_p, a 32 point approximation of the associated acoustic spectral envelope is produced. Then if the desired value for the *i -th* component of the spectrum is t_{ip}, and the approximated value is o_{ip}, we can define an error measure for that approximate envelope to be,

$$E_p = \sum_{1}^{32} (t_{ip} - o_{ip})^2 \max(t_{ip}, o_{ip})$$

This weighting scheme reflects the relative importance of the the height of spectral envelope as demonstrated by speech recognition tasks (Miller, 1953). This is also the error measure used by many speech recognition systems to compare an unknown acoustic spectra to a set of stored templates (Klatt, 1976; Jelinek, 1985).

The second criterion used to evaluate the network is a forced choice test. In this test, the spectral estimate obtained from a given image is compared to a set of known spectral envelopes. Based on the prior error measure, the closest envelope is identified and selected. A selection is considered successful if it is the actual spectrum associated with that original image or if it is spectrum associated with another example of the same vowel. In addition to the number of correct matches, the confusions are also of interest since they reveal whether the network is processing the images in a manner similar to humans lip readers. As part of this research, the networks confusions will be compared to those confusions made by other estimation methods, and by humans lip readers.

RESULTS AND EVALUATION

Over the last year, we have constructed and trained numerous neural networks. These networks have allowed us to explore the effects of the connectivity structure, the number of hidden units, and the type of transfer function used by the processors. In addition to the comparisons amongst the different networks, the best network estimates were then compared to a variety of other estimation methods.

While it was difficult to assess these estimates from the value of the error function alone, the forced choice test has provided a tangible measure of performance. When compared to the performance of humans on the similar task of classifying vowels from the visual signal alone, the networks performed better. The acoustic envelope estimated by the networks was correctly matched to the same vowel from 62% to 68% of the time. In comparison, human lipreaders have demonstrated in previous research performance levels of between 49% and 54% (Jackson, et.al., 1986; Berger, 1970;

Montgomery and Jackson, 1983; Erber, 1979).

The networks' estimations were also compared with estimates obtained from a template matching approach used extensively for pattern matching. Those images used to train the network were now used as the stored templates. Given a new image, the closest match or matches were found among the stored images and the corresponding spectral envelopes were averaged together. This is then used as the estimate of the spectrum envelope associated with that image.

The quality of the estimate depended heavily on how many spectral envelopes were averaged. It was found that finding the five closest images provided the best estimates. Using five envelopes, the template matching method performed on par with the neural networks.

This alone would not be remarkable, unless one considers the problems involved in implementating the template method. As the number of templates increases, the computation needed to do the comparison and ranking increases as $O(N^2)$, where N is the number of pixels in the images. The only solution is to create a smaller set of templates. However, Petajan's work (1988) showed that encoding the images resulted in poorer performance for a similar lip reading tasks.

WORK IN PROGRESS

In addition to the template matching method, the networks are being compared to other estimation techniques. Most estimation procedures are designed to work with input and output data of small dimensionality. As a result of this, it is often necessary to encode the input and output data before some type of regression is attempted. The problem is choosing the correct encoding. If the parameters in the input that are necessary to predict the output are not clearly identified, then one has two choices.

First, one can select some parameters a priori, and test to see how well the parameters correlate with the output, and account for the variance. Towards this end, we will review those parametric studies of human lip reading that have been performed (Montgomery and Jackson, 1983). The goal of these studies was to determine those parameters available in the visual signal that could determine the vowel being spoken. The comparison will allow us to evaluate how well the network is selecting its parameters as compared to *experts*.

The second method is to choose some encoding based upon some known criteria, such as linear least-square-error (LLSE) encoding (Gonzales and Wintz, 1977). The optimal LLSE encoding can be obtained using principal component analysis. Using this encoding the images and their associated acoustic spectra will be represented in terms of their principal components. Next, an attempt will be made to fit a linear mapping from the input data set to the output data set using linear regression. Once a fit is defined, a new image could be encoded and used to construct an estimate of the associated spectra. This estimate will be in terms of the principal components used to describe the acoustic envelopes in the training set. This method will reveal whether or not an optimal LLSE encoding will collapse the data along dimensions which are vital to the problem under study. There is no reason to believe that it won't.

One of the benefits of using neural networks is the ease with which additional constraints can be introduced. In the coming year, we hope to improve the performance of the networks by using the dynamical constraints inherent in the speech production process. As part of this effort, we intend to use networks with recurrent links that will allow us to work with sequences of images.

LONG TERM IMPLICATIONS

The approach taken by this research may provide a basis for a new generation of speech recognition systems that use two sensory channels. In designing such a system, the engineer can benefit by looking at the best speech recognition system around, the human being. The parts of this system that are the most fully studied and best understood are the human acoustic and visual preprocessing systems. Already, acoustic speech recognition systems are benefitting from what is known about the human auditory system by using models of the human ear as a front end (Jelinek, F., 1985).

At Caltech, Carver Mead has already successfully designed and fabricated a variety of synthetic retinas and cochleas in analog VLSI. These peripheral systems process massive amounts of sensory

data in real time, and output a distilled, parallel and analog representation. Given these parallel output from two channels, the question is how to combine them.

The traditional approach would be to encode their outputs symbolically and to try and define constraints between these two symbol strings. One of the problems with encoding these signals from these two channels is that the symbolic encoding can obscure constraints that might otherwise be useful, or quite simply might throw the information away.

The alternative approach is to maintain the distributive representation that comes out of these channels and attempt to combine them at a sub-symbolic level. This research looks at the feasibility of this second approach.

Jelinek, F., "The Development of an Experimental Discrete Dictation Recognizer," *Proceedings IEEE*, vol. 73, pp. 1616-1624, 1985. The Special Issue on Man-machine Speech Communication.

Klatt, D.H., *A Digital Filter Bank for Spectral Matching*, Philadelphia, PA, April 1976. Proceedings of ICASSP 1976

Klatt, D.H., "Software for a cascade/parallel formant synthesizer," *JASA*, vol. 67, pp. 971-995 , March 1980 .

Kohonen, T., *Self-Organization and Associative Memory*, Springer-Verlag, Berlin, 1984.

Lippmann, R.P. and Gold, B., *Neural-Net Classifiers Useful for Speech Recognition*, June 1987 . presented at Internation Conf on Neural Networks in San Diego, CA

Lippman, R.P., "An Introduction to Neural Networks," *IEEE ASSP MAGAZINE*, pp. 4-22, April 1987.

Marr, D. and Poggio, T., "Cooperative computation of stereo disparity," *Science*, vol. 194, pp. 283-287, 1976.

Poggio, T., Torre, V., and Koch, C., "Computational Vision and regularization theory," *Nature*, vol. 317, pp. 314-319, 1985.

McGurk, H. and MacDonald, J., "Hearing Lips and Seeing Voices," *Nature*, vol. 264, pp. 746-748, 1976.

McGurk, H. and MacDonald, J., "Visual influences on speech processes," *Perception & Psychophysics*, vol. 24, pp. 253-257, 1978.

Miller, G.A. and Nicely, P.E., "An analysis of perceptual confusions among some English consonants," *Journal of the Acoustical Society of America*, vol. 27, pp. 338-352, 1955.

Miller, R.L., "Auditory Tests with Synthetic Vowels," *JASA*, vol. 25, pp. 114-121, January 1953.

Minsky, M. and Papert, S., *Perceptrons: An Introduction to Computational Geometry*, The MIT Press, Cambridge, MA, 1972.

Montgomery, A. and Jackson, P.L., "Physical Characteristics of the lips underlying vowel lipreading," *Journal of the Acoustical Society of America*, vol. 73, pp. 2134-2144, 1983.

Nishida, Shogo, "Speech Recognition Enhancement by Lip-Information," *CHI'86 Conference Proceedings*, pp. 198-204, ACM, April, 1986. Special Issue of the SIGCHI Bulletin

Oppenheim, A.V. and Schafer, R.W., *Digital Signal Processing*, Prentice-Hall, Inc, Englewood Cliffs, NJ, 1975.

Peeling, S.M. and Bridle, J.S., *Experiments with a Learning Network for a Simple Phonetic Recognition Task*, 1987 . Proc. IoA Autumn Conf of Speech and Hearing

Peeling, S.M., Moore, R.K., and Tomlinson, M.J., *The Multi-Layer Perceptron as a Tool for Speech Pattern Processing Research*, 1987 . Proc. IoA Autumn Conf of Speech and Hearing

Petajan, E.D., *Automatic Lipreading to Enhance Speech Recognition*, pp. 265-272, IEEE Communication Society, November 26-29,1984. Global Telecommunications Conference

Petajan, E.D,, "An improved Automatic Lipreading System To Enhance Speech Recognition," *Bell Laboratories Technical Report*, no. 11251-871012-111TM, 1987.

Peterson, G.E. and Barney, H.L., "Control Methods Used in a Study of the Vowels," *J. Acoust. Soc. Am*, vol. 24, pp. 175-184, March 1952.

Pickett, J.M., *The Sounds of Speech Communication: A Primer of Acoustic Phonetics and Speech Perception*, University Park Press, Baltimore, MD, 1980.

Pineda, F.J., "Generalization of Back-Propagation to Recurrent Neural Networks," *Physical Review Letters*, vol. 59, pp. 229-2232 , 1987.

Prager, R.W., Harrison, T.D., and Fallside, F., *Boltzmann Machines for Speech Recognition*, Cambridge University, 1986. Technical Report CUED/F-CAMS/TR.260

Rabiner, L.R. and Schafer, R.W., *Digital Processing of Speech Signals*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1978.

Reddy, D.R., "Segmentation of Speech Sounds," *JASA*, vol. 40, pp. 307 312, April 1966.

Reddy, D.R., "Computer Recognition of Connected Speech," *JASA*, vol. 42, pp. 329-347, April 1967.

Rumelhart, D.E., Hinton, G.E., and Williams, R.J., *Learning internal representations by error propagation*, 1, MIT Press, Cambridge, MA, 1985. Parallel Distributed Processing in the Microstructure of Cognition: Vol 1. Foundations

Rumelhart, D.E., Hinton, G.E., and McClelland, J.L., "A General Framework for Parallel Distributed Processing," in *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, ed. McClelland, J.L, MIT PRESS, 1987.

Rumelhart, D.E., McClelland, J.L., and the PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, I, The MIT Press, Cambridge, Mass, 1986.

Sejnowski, T.J. and Rosenberg, C.R., *NETtalk: A parallel network that learns to read aloud*, 1986. Johns Hopkins University Department of Electrical Engineering and Computer Science Technical Report 86/01

Sejnowski, T.J. and Rosenberg, C.R., "Parallel Networks that Learn to Pronounce English Text," *Complex Systems*, vol. 1, pp. 145-168, 1987.

Sejnowski, T.J. and Hinton, G.E., "Parallel Stochastic Search in Early Vision," in *Vision, Brain and Cooperative Computation*, ed. Hanson, A.R., 1984.

Sejnowski, T.S., Keinker, P., and Hinton, G.E., "Learning Symmetry Groups with Hidden Units: Beyond the Perceptron," *Physica D.*, 1985.

Stevens, K.N. and House, A.S., "Development of a Quantitative Description of Vowel Articulation," *JASA*, vol. 27, pp. 484-493, 1955.

Sumby, W.H. and Pollack, I., "Visual Contribution to Speech Intelligibility in Noise," *The Journal of the Acoustic Society of America*, vol. 26, pp. 212-215, March 1954.

Summerfield, Q., "Audio-visual Speech Perception, Lipreading and Artificial Stimulation," in *Hearing Science and Hearing Disorders*, ed. Haggard, M.P., pp. 131-182 , Academic Press, London, 1982.

Summerfield, Quentin, "Use of Visual Information for Phonetic Perception," *Phonetica*, vol. 36, pp. 3'4-331, 1979.

Walden, B.E., Prosek, R.A., Montgomery, A., Scherr, C.K., and Jones, J.J., "Effects of Training on the Visual Recognition of Consonants," *Journal of Speech and Hearing Research*, vol. 20, pp. 130-145, 1977.